Advancing Isolated Saudi Sign Language Recognition Using Transformer models

1st SOUKEINA ELHASSEN
dept. Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
selhassen@stu.kau.edu.sa

2nd Lama Al Khuzayem
dept. Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
lalkhuzayem@kau.edu.sa

3rd Areej Alhothali
dept. Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
aalhothali@kau.edu.sa

4th Ohoud Alzamzami
dept. Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
ualzamzami@kau.edu.sa

5th Nahed Alowaidi
dept. Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
nalowidi@kau.edu.sa

***Abstract*—Sign Language (SL) is the primary communication method for deaf and hard-of-hearing individuals, which requires advanced technologies that facilitate the communication between SL users and hearing community. Saudi Sign Language (SSL) is the main SL language at Saudi Arabia, and suffers from the lack of large isolated datasets, requiring solutions that provide good results on small to medium datasets. We propose in this work a benchmark for isolated SSL recognition from transformer-based video models , we study the effectiveness of VideoMAE in details at the SL recognition with comparison to ViViT, and TimeSformer, the models are fine-tuned on the KSU-SSL dataset with 15,999 videos spanning 80 signs. The models, are pre-trained on Kinetics-400 dataset, with 16-frame RGB clips and classify signs to achieve robust spatiotemporal feature extraction. VideoMAE achieved the optimal results at 95.25% accuracy, followed by TimeSformer (93.44%) and ViViT (92.81%). Despite challenges with visually similar signs, the benchmark enhances accessibility for Saudi Arabia's deaf community, enabling real-time translation and pedagogical support. This research serves as a benchmark for Saudi SLR, with opportunities for continued SSL and cross-lingual Arabic sign language recognition.**

***Index Terms*—Arabic Sign Language, Isolated Sign Language, Saudi Sign Language, Sign Language Recognition, Transfer Learning, VideoMAE, ViViT, TimeSformer, KSU-SSL Dataset, Video Transformers, Deep Learning**

I. Introduction

Sign language is the primary method for deaf and hard-of-hearing community to communicate through the world, giving it a global significance. Saudi Sign Language (SSL) is the primary language for the deaf and hard-of-hearing community in Saudi Arabia, it features a unique cultural vocabulary specific to SSL and some shared vocabulary with Arabic Sign Language (ArSL). Saudi Sign Language has inherited the linguistic difficulties inherent in Arabic, including the multiplicity of terms for a single word and sentence structures, making it an important area of research. The statistics estimate approximately 720,000 SSL users in Saudi Arabia[1], highlighting the need for advanced technologies that help the deaf community to have better communication with hearing society, allowing them to acquire appropriate services in education, healthcare, and social interactions. Despite the importance of Saudi Sign Language, there is limited research addressing it compared to ArSL and other sign languages which lead to the scarcity of SSL datasets. Most of the available dataset is Image-based dataset which lead to developing models that suffer from the difficulty of capturing dynamic movements in video. Most solutions focus on traditional convolutional neural network (CNN)-based approaches, which suffer from the difficulty of capturing movements that extend over a period of time. In this research we introduce the use of transformers based models as a promising solution for capturing global temporal dependencies and better video understanding [2].

This study proposes fine-tuning three pre-trained video transformer models—VideoMAE [3], ViViT [4], and TimeSformer [5] on the KSU-SSL dataset[1] for the recognition of isolated SSL. These models leverage self-attention mechanisms to capture spatiotemporal features effectively, which consider crucial for modeling dynamic hand and body gestures in SSL videos. VideoMAE's masked autoencoding enhances feature robustness, ViViT's factorized attention balances spatial and temporal modeling, and TimeSformer's divided space-time attention optimizes efficiency. Our approach addresses the low resource nature of SSL data, enabling robust performance despite limited training samples, by utilizing transfer learning with pre-trained weights from large-scale video datasets. The study demonstrated the effectiveness of using transfer learning, achieving standard performance across all systems with variations between them, which helps in identifying which features in the transformer systems are effective in learning sign language. In this work we aim to: (1) evaluate the accuracy and robustness of VideoMAE, ViViT, and TimeSformer for isolated SSL recognition; (2) establish a comprehensive bench-

Identify applicable funding agency here. If none, delete this.

mark for the KSU-SSL dataset to guide future SSL research; and (3) provide insights into the efficacy of transformer-based models for sign language tasks. Our paper addresses a comparative analysis of the three models, highlighting their strength, and weakness, and implications on how to promote Saudi Arabia's deaf community accessibility. These developments design superior communication devices and promote inclusive technology for low-resource sign languages.

The paper is organized as follows: Section 2 reviews related work on sign language recognition. Section 3 details the methodology, including dataset and model descriptions. Section 4 presents experiments and results. Section 5 discusses findings, and Section 6 concludes with future research directions.

II. Related Work

Sign language recognition (SLR) has evolved a lot from the early hand-tuned feature methods such as histogram of oriented gradients to today's sophisticated deep learning methods [6]. In the past, SLR systems relied on time-consuming manual feature engineering that did not handle gesture and lighting variations very well. Deep learning and convolutional neural networks (CNN) revolutionized SLR entirely by offering methods to learn features directly from video streams. SLR is categorized as isolated SLR, with word-level classification, and continuous SLR, with sentence-level sequences. Isolated SLR is the focus of this work because it minimizes the complexity in gesture classification but remains challenging depending on different gestures. CNN-based models, such as 3D-CNNs, were most notable in early deep learning SLR applications, achieving good results on datasets such as ASL Lexicon. Yet, CNNs tend to miss capturing long-term temporal dependencies essential for dynamic signs, leading to investigation into state-of-the-art architectures such as transformers for enhanced spatiotemporal modeling within SLR tasks.

ArSL and SSL recognition studies have been under growing attention, due the appearance of datasets like KArSL and KSU-SSL [7]. KArSL, a heterogeneous ArSL dataset with 502 signs from multiple signers, acts as the foundation of gesture recognition studies [7]. KSU-SSL, in the context of SSL, offers a unique collection of isolated signs, capturing regional linguistic characteristics. Experiments on ArSL/SSL tend to employ CNN-based models or CNN-RNN hybrid models, which work reasonably well in recognizing static and dynamic gestures. For instance, CNN models employed on KArSL worked well for isolated signs but did a poor job with the temporal relationships in complex gestures. The limitations indicate the need for additional models that have better long-range dynamic capture. KSU-SSL's focus on SSL-specific features renders it a necessary tool, yet its applicability is not yet realized since not many have utilized higher-end architectures, necessitating innovative approaches in SSL detection. The shift in paradigm to transformer-based models has transformed SLR, employing self-attention mechanisms to model spatiotemporal dependencies effectively. Studies utilize transformer models for datasets like WLASL, a large-scale ASL dataset, and BdSLW60, a Bangla sign dataset, have shown improved performance over CNNs with accuracy levels higher than 90% for isolated signs [8]. The successes are due to the ability of transformers to model long-range dependencies, crucial for dynamic gestures. However, their application on SSL remains scarce, which motivates us to explore VideoMAE, ViViT, and TimeSformer to advance SSL recognition and confirm their utility under low-resource settings. Transfer learning is now being at the core of SLR, addressing low dataset sizes in low-resource sign languages like SSL. Large-scale video dataset pre-trained models like Kinetics-400 come with robust initial weights enabling fine-tuning on low-sized SLR datasets. This is enabled by strong model generalization, particularly for isolated sign recognition where data sparsity is typically the norm. Fine-tuning CNNs on WLASL for instance shows that it improve accuracy by leveraging pre-trained features [8]. Transfer learning has limited use in SSL recognition with little work conducted on pre-trained transformer models, while it show successful results in ASL and other sign languages . This research aims to bridge the gap by employing transfer learning through VideoMAE, ViViT, and TimeSformer on the KSU-SSL dataset.

III. Methodology

This study fine-tunes three video transformer models (VideoMAE [3], ViViT [4], and TimeSformer [5]) on the KSU-SSL dataset for isolated SSL recognition. The isolated SSL recognition task is framed as a video classification task using a pre-trained VideoMAE model. Each video input is a sequence of frames $x = \{x_1, x_2, \ldots, x_{16}\}$, where each frame $x_i \in \mathbb{R}^{224\times224\times3}$ is an RGB image of dimension 224×224. The model takes in 16 frames from one video clip, which are uniformly sampled from the video length. Frames are resized and normalized according to mean and standard deviation values specified in the preprocessor configuration. Target output y is a class label from a list of 80 predefined classes with each label representing an isolated SSL gesture.

The overall objective is to learn a function f which accepts the input video sequence and transforms it into a probability distribution over the 80 class labels. The model learns spatial features and temporal features from the video frames based on its architecture being a transformer-based one, yielding an output representation that passes through a linear classifier. The classifier, using a softmax function, approximates the probability of each class. The final prediction is achieved by taking the class with highest probability:

$$\hat{y} = \arg \max_{j \in \{0,\ldots,99\}} p(y_j \mid x)$$

It is trained using cross-entropy loss, which approximates the difference between the estimated class probabilities and the true one:

$$\mathcal{L} = -\sum_{j=0}^{99} y_j \log(p(y_j \mid x))$$

It trains using transfer learning, where the initial layers of the model for general feature extraction are frozen so that pre-training gained knowledge is retained. The final classification layers alone are fine-tuned on the KSU-SSL dataset such that the model learns to generalize to the new features of the data and also retain good video representation capabilities. Fine-tuning is achieved by calculating the loss of every prediction and adjusting the model parameters in order to minimize the loss, aiming at having precise isolated SSL gesture classification. This approach is a trade-off between pre-training and general large-scale knowledge and the specific requirements of SSL sign recognition.

The approach can learn spatial and temporal features required for gesture recognition with strong performance on low-resource SSL data. This benchmark will open the door towards better access for the deaf population in Saudi Arabia by giving a basis for automatic recognition of SSL.

A. KSU-SSL Dataset

KSU-SSL dataset is comprises of 15,999 videos for 80 single SSL signs signed by 40 signers captured by cameras in RGB mode with uniform lighting, backgrounds, and signing styles [1]. Videos are static and dynamic signs of varied duration, videos were normalized to 30 FPS and resized to 224×224 pixels during preprocessing. Random crop, horizontal flip, and temporal subsampling are used as augmentations for improved robustness. The dataset is allocated 80% train (12,799 videos), 10% validation (1,600 videos), and 10% test (1,600 videos). KSU-SSL's Saudi-centric signs make it an inherent fit for isolated SLR, although complexity like signer variability and occlusion requires careful preprocessing. This arrangement allows generalization to new signers, a real-world SSL task requirement.

The VideoMAE model is initially pre-trained on the Kinetics-400 dataset using self-supervised learning to develop robust spatiotemporal representations. In our work, transfer learning was employed by fine-tuning the model on the KSU-SSL dataset to adapt it for isolated SSL recognition tasks. The Kinetics-400 dataset comprises 400 action classes, each containing 400 video clips. The utilization of pre-trained self-supervised learning has demonstrated its reliability, as evidenced in this study [10]. The use of transfer learning helps capture general video features, making it well-suited for tasks involving complex temporal dynamics. The training process included careful data preprocessing, involving temporal subsampling, resizing, normalization, and augmentations like random cropping and horizontal flipping. These transformations introduced variability and enhanced model generalization. The model was trained with a learning rate of 3e-5, a batch size of 4, and a warmup ratio of 0.1, ensuring a balanced learning process. Additionally, early stopping was implemented to avoid overfitting. The fine-tuning process used a structured dataset split into training, validation, and testing subsets, with specific transformations applied to each video to meet the model's input requirements.

In addition to the primary VideoMAE approach, we used transfer learning with the ViViT (Video Vision Transformer) model and TimeSformer to assess the KSU-SSL dataset's performance. The efficacy of the transformer-based model is demonstrated in spatiotemporal video recognition tasks. ViViT [4] was created to expand the functionality of transformers, initially developed for image processing, to include video data. It is a flexible option for modeling intricate motion and gesture-based interactions, as it utilizes spatial and temporal tokens in a highly modular structure. By evaluating this model on the KSU-SSL dataset, we aim to investigate the dataset's effectiveness and compatibility across various architectures and gain a better understanding of how it might improve the sign language recognition field.

B. Pretraining Phase

The VideoMAE model is pretrained on the Kinetics-400 dataset [3], which contains approximately 240,000 video clips covering 400 human actions, with most clips lasting around 10 seconds. Pretraining employs a self-supervised masked autoencoding approach to develop generalizable video representations without labeled data.

a) Input Representation: Each video is represented as a sequence of frames $x = \{x_1, x_2, \ldots, x_T\}$, where each frame $x_i \in \mathbb{R}^{224\times224\times3}$ is an RGB image resized to 224×224 pixels with 3 channels. Videos are processed into a tensor $V \in \mathbb{R}^{T\times H\times W\times C}$, where $T = 16$, $H = W = 224$, and $C = 3$. Frames are normalized using ImageNet statistics:

$$\begin{aligned} V_{\text{norm}} &= \frac{V - \mu}{\sigma}, \\ \mu &= [0.485, 0.456, 0.406], \\ \sigma &= [0.229, 0.224, 0.225] \end{aligned} \tag{1}$$

Each frame is divided into non-overlapping patches of size $P \times P = 16 \times 16$, yielding $N_f = \frac{224}{16} \times \frac{224}{16} = 196$ patches per frame and $N = 16 \times 196 = 3136$ total patches. Patches are flattened and embedded:

$$x_p = \text{Flatten}(V_{t,i,j}) \in \mathbb{R}^{P^2\cdot C}, \quad x_e = W_e x_p + e_{\text{pos}} \in \mathbb{R}^d \tag{2}$$

where $W_e \in \mathbb{R}^{d\times(P^2\cdot C)}$ is the embedding matrix, $d = 768$, and $e_{\text{pos}} \in \mathbb{R}^{N\times d}$ encodes spatial and temporal positions.

b) Masked Autoencoding: During pretraining, approximately 75% of patches are randomly masked, setting their pixel values to zero ($m = 0.75$). Let $M \subset \{1, \ldots, N\}$ be the indices of unmasked patches, with $|M| = (1 - m) \cdot N$. The input tokens are:

$$x_m[i] = \begin{cases} x_e[i], & \text{if } i \in M \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The model reconstructs masked patches, minimizing the mean squared error (MSE) loss:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|M^c|} \sum_{i\in M^c} \|V[i] - \hat{V}[i]\|_2^2 \tag{4}$$

where $M^c = \{1, \ldots, N\} \setminus M$ contains masked patch indices, and $\hat{V}$ is the reconstructed video. This process enhances the model's ability to learn spatiotemporal dependencies, making it robust to noise and occlusions in SSL videos.

c) Model Architecture: The VideoMAE architecture consists of an encoder and a decoder, both transformer-based.

d) Encoder: The core of the model is the transformer network, where the encoder receives the embedded patches, capturing contextual relationships between patches using multi-head self-attention mechanisms. The transformer consists of multiple layers including multi-head self-attention, feed-forward neural network, and positional encoding. The encoder processes unmasked patch tokens $X \in \mathbb{R}^{|M| \times d}$ through $L = 12$ layers, each including:

- **Multi-Head Self-Attention (MHSA)**:

$$Q_h = XW_{Q_h}, \quad K_h = XW_{K_h}, \quad V_h = XW_{V_h} \tag{5}$$

$$A_h = \text{Softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_h}}\right), \quad \text{Head}_h = A_h V_h \tag{6}$$

$$\text{MHSA}(X) = \text{Concat}(\text{Head}_1, \ldots, \text{Head}_H) W_O \tag{7}$$

  where $W_{Q_h}, W_{K_h}, W_{V_h} \in \mathbb{R}^{d \times d_h}$, $H = 12$, $d_h = d/H = 64$, and $W_O \in \mathbb{R}^{d \times d}$.
- **Feed-Forward Network (FFN)**:

$$\text{FFN}(x) = W_2 \cdot \text{GELU}(W_1 x + b_1) + b_2 \tag{8}$$

  where $W_1 \in \mathbb{R}^{d_{\text{ff}} \times d}$, $W_2 \in \mathbb{R}^{d \times d_{\text{ff}}}$, and $d_{\text{ff}} = 3072$.
- **Layer Normalization**:

$$\text{LayerNorm}(x) = \frac{x - \mu_x}{\sigma_x} \cdot \gamma + \beta \tag{9}$$

The layer computation is:

$$X' = \text{MHSA}(\text{LayerNorm}(X)) + X \tag{10}$$

$$X'' = \text{FFN}(\text{LayerNorm}(X')) + X' \tag{11}$$

e) Decoder: In the pretraining phase, the model uses a lightweight decoder to reconstruct video frames using the encoded representations, forcing the model to learn relationships between masked and non-masked patches. The encoder outputs $Z \in \mathbb{R}^{|M| \times d}$ are combined with mask tokens:

$$Z_{\text{dec}}[i] = \begin{cases} Z[i], & \text{if } i \in M \\ m_{\text{token}}, & \text{if } i \in M^c \end{cases} \tag{12}$$

A lighter transformer (8 layers, $d_{\text{dec}} = 512$) processes Z_{dec}, and a linear head reconstructs patches:

$$\hat{V}_i = W_r Y_i + b_r \tag{13}$$

where $W_r \in \mathbb{R}^{(P^2 \cdot C) \times d_{\text{dec}}}$. This process enhances robustness to noise and occlusions, critical for video tasks like SSL recognition. The decoder is not part of the fine-tuning pipeline, where the model is fine-tuned for classification using VideoMAEForVideoClassification, which replaces the decoder with a classification head.

C. Fine-Tuning Phase

The pretrained VideoMAE model is fine-tuned on the KSU-SSL dataset, which contains 15999 training samples, to classify 80 isolated SSL signs. The model architecture is illustrated in Fig. 1.

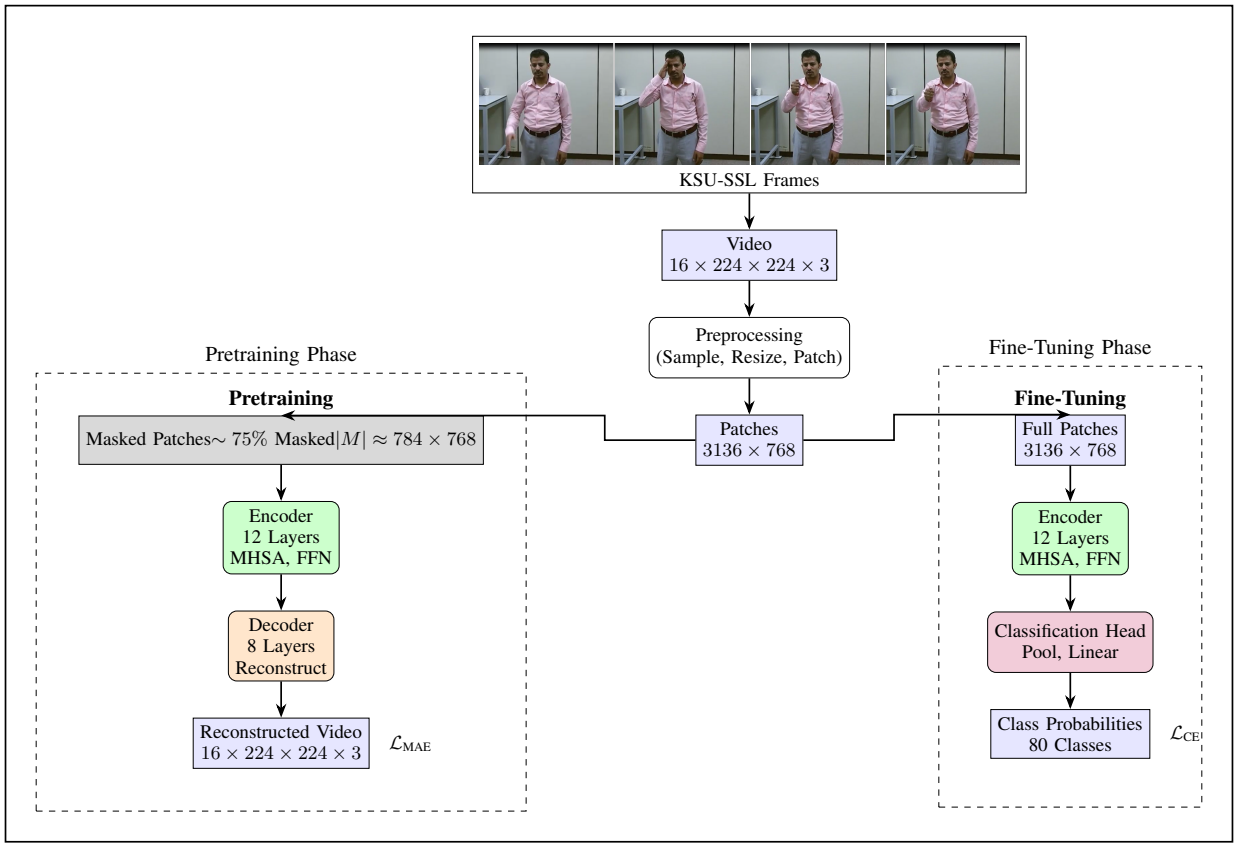

Fig. 1: Model Architecture for Isolated SSL Recognition. The pipeline starts with KSU-SSL dataset frames, followed by preprocessing. *Pretraining* uses masked autoencoding to reconstruct videos, while *fine-tuning* classifies 80 isolated SSL gestures.

a) Input Representation and Data Preprocessing: The model was fine-tuned on the KSU-SSL dataset to achieve the task of isolated SSL recognition. The dataset was split into three subsets: 80% training, 10% validation, and 10% testing.

The dataset has been modified by adding padding to video frames since our dataset is captured using mobile cameras, resulting in rectangular videos, while the model accepts square videos. After padding, the input frames are processed into a sequence of non-overlapping patches 16×16 pixels, resulting in 196 patches per frame and 3,136 patches across 16 frames. This process is similar to Vision Transformers (ViTs) for handling images. These patches are then flattened and embedded to preserve their order. Table **??** illustrates a sample of ground truth data.

Videos are processed into a tensor $V \in \mathbb{R}^{T \times H \times W \times C}$, where $T = 16$ frames, $H = W = 224$ pixels (after resizing), and $C = 3$ (RGB channels). Frames are normalized using ImageNet statistics:

$$\begin{aligned} V_{\text{norm}} &= \frac{V - \mu}{\sigma}, \\ \mu &= [0.485, 0.456, 0.406], \\ \sigma &= [0.229, 0.224, 0.225] \end{aligned} \tag{14}$$

Each frame is divided into non-overlapping patches of size $P \times P = 16 \times 16$, yielding $N_f = \frac{224}{16} \times \frac{224}{16} = 196$ patches per frame and $N = 16 \times 196 = 3136$ total patches. Patches are flattened and embedded:

$$x_p = \text{Flatten}(V_{t,i,j}) \in \mathbb{R}^{P^2 \cdot C}, \quad x_e = W_e x_p + e_{\text{pos}} \in \mathbb{R}^d \tag{15}$$

where $W_e \in \mathbb{R}^{d \times (P^2 \cdot C)}$ is the embedding matrix, and $e_{\text{pos}} \in \mathbb{R}^{N \times d}$ encodes spatial and temporal positions. The preprocessing includes augmentations such as random cropping, horizontal flipping, and uniform temporal subsampling.

The dataset frames were resized to 224×224, and augmentation techniques were applied, such as random crop, random horizontal flip, and random short side scale. The videos were labeled with 80 distinct isolated SSL signs. The original dataset consists of raw videos representing various isolated SSL signs, varying in resolution, lighting conditions, backgrounds, and signing styles, introducing natural diversity to the data. However, this variability can sometimes hinder the model's ability to generalize across different conditions. To address this, a series of augmentation techniques were applied to improve the model's robustness.

The dataset videos, after augmentation, are transformed to include variations that simulate different real-world conditions. The augmentations applied include random cropping, resizing to a fixed dimension of 224×224, horizontal flipping with a probability of 0.5, and temporal subsampling to ensure consistent frame sampling across clips. Additionally, normalization is applied using predefined mean and standard deviation values. These augmentations help the model learn more generalized features by exposing it to a wider range of visual variations, ultimately improving its performance on unseen data. The augmented dataset ensures that the model can better handle diverse video inputs, leading to improved classification accuracy and robustness during testing.

TABLE I: Sample Ground Truth Labels for KSU-SSL Dataset

ID	Class Label	Type
1	Hello	Dynamic
2	Thank you	Dynamic
3	Mother	Dynamic
4	Father	Dynamic
5	Translator	Dynamic
6	Deafness	Dynamic
7	Help	Dynamic
8	Food	Dynamic
9	Water	Dynamic
10	1	Static
11	2	Static
12	A	Static
13	B	Static

b) Model Architecture: For fine-tuning, we replace the pretraining decoder with a classification head. The encoder (same as pretraining) processes all patches ($X \in \mathbb{R}^{N \times d}$, $N = 3136$) without masking. Outputs are pooled:

$$z_{\text{pool}} = \frac{1}{N} \sum_{i=1}^{N} Z_i \tag{16}$$

A linear layer produces logits for $K = 100$ classes:

$$\text{logits} = W_c z_{\text{pool}} + b_c, \quad W_c \in \mathbb{R}^{K \times d} \tag{17}$$

The objective is to learn a function f mapping videos to class probabilities:

$$\hat{y} = \arg \max_{j \in \{0,\ldots,99\}} p(y_j \mid x) \tag{18}$$

The model is trained with cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{j=0}^{99} y_j \log(\text{Softmax}(\text{logits})_j) \tag{19}$$

c) Training Setup: Fine-tuning is performed on the Aziz Supercomputer, utilizing its resources to train the complex transformer model leveraging parallel GPU processing. Training parameters are:

- **Epochs**: 50, with early stopping after 3 epochs of no improvement (stopped at 8 epochs).
- **Learning Rate**: 3e-5.
- **Batch Size**: 4.
- **Weight Decay**: 0.01.
- **Warmup Ratio**: 0.1.

Videos are loaded using a custom dataset class, sampling 16 frames uniformly per video via linear interpolation and applying transformations (resizing, cropping, flipping, normalization).

d) Model Optimization: To prevent overfitting, a weight decay of 0.01 regularizes the model. The learning rate of 3×10^{-5} balances convergence speed and stability, while a warmup ratio of 0.1 ensures gradual learning rate increases early in training.

e) Why VideoMAE for SSL?: VideoMAE is a suitable option for isolated SSL recognition due to several reasons:*Generalization Across Domains:* The Kinetics-400 dataset, on which VideoMAE was pre-trained, is relevant to SSL recognition as a subset of action and gesture recognition, enabling efficient adaptation to isolated SSL signs. Pre-trained on large-scale datasets, VideoMAE captures general motion patterns viable across domains, including sign language. When fine-tuned on KSU-SSL, these representations adapt to SSL-specific nuances, enhancing generalization. *Handling Temporal Dependencies:* VideoMAE captures spatial and temporal dependencies crucial for isolated SSL, where individual signs rely on hand and body movements. Its ability to model short-term dependencies ensures accurate recognition of distinct gestures, adapting to variations in signing speed and style. *Enhance Generalization and Robustness:* The masked autoencoder approach improves model generalization by predicting missing patches, enhancing robustness to noise and signer variations in KSU-SSL videos. This reconstruction capability results in a more robust model for isolated signs. *Effective Feature Representation:* VideoMAE addresses the scarcity of large-scale annotated SSL datasets by leveraging self-supervised learning to capture robust spatiotemporal features. It captures local patterns (e.g., hand shapes) and global patterns (e.g., gesture context), essential for isolated SSL recognition. *Noise Robustness:* VideoMAE's masked pre-training enhances resilience to noise and occlusions, common in real-world SSL scenarios (e.g., varying camera angles, partial hand visibility), ensuring reliable performance on KSU-SSL.

f) Exploration with ViViT and TimeSformer: The KSU-SSL dataset was validated by ViViT (Video Vision Transformer) model [4] and TimeSformer (Time-Space Transformer)[5] to assess SSL compatibility with other architectures. We fine-tune both on KSU-SSL for isolated SSL recognition, following a methodology similar to VideoMAE. Both models leverage self-attention mechanisms to capture spatiotemporal features, critical for modeling SSL's dynamic gestures, and are pre-trained on Kinetics-400 [9] to ensure robust initial representations. ViViT extends Vision Transformers to videos by factorizing spatial and temporal attention, enabling modular gesture modeling. Videos are processed as a sequence of frames $x = \{x_1, \ldots, x_{16}\}$, where each frame $x_i \in \mathbb{R}^{224\times224\times3}$ is an RGB image. Frames are divided into 16×16 patches, yielding $N_f = 196$ patches per frame and $N = 16 \times 196 = 3136$ total patches.

TimeSformer suggests divide space-time self-attention, where spatial and temporal dimensions are attended to sequentially for efficiency. Videos are also represented as sequences of 16 frames, patches embedded similarly to ViViT. TimeSformer attention computes spatial attention frame-wise, then temporal attention for each spatial location across frames:

$$A_{\text{spatial}} = \text{Softmax}\left(\frac{Q_s K_s^T}{\sqrt{d_h}}\right), \quad A_{\text{temp}} = \text{Softmax}\left(\frac{Q_t K_t^T}{\sqrt{d_h}}\right) \tag{20}$$

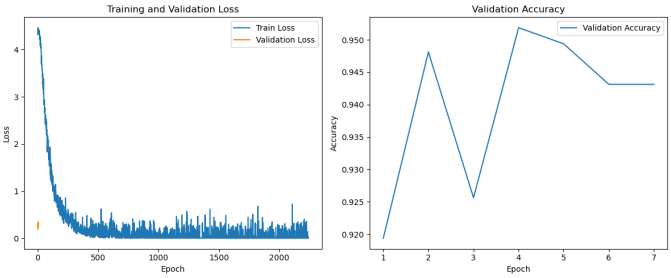

Fig. 2: Training and Validation loss

where $Q_s, K_s, Q_t, K_t \in \mathbb{R}^{d_h}$, and $d_h = 64$. The architecture has 12 layers, 12 heads, and a hidden dimension of 768. Supervised classification on Kinetics-400 is pre-trained using cross-entropy loss. This approach provides higher importance to long-range dependencies, which is crucial for SSL's dynamic gestures. TimeSformer is fine-tuned on KSU-SSL with the same pre-processing as VideoMAE and ViViT, normalization, and augmentations. 80-class logits are output from the classification head during training with cross-entropy loss and the same hyperparameters (learning rate: 3e-5, batch size: 4, weight decay: 0.01, warmup ratio: 0.1).

IV. Results

The models were evaluated using metrics such as overall accuracy, precision, recall, confusion matrix, and class-wise performance on the test subsets of KSU-SSL, which provides insights into strengths and weaknesses. A validation accuracy of 95.25% was achieved by VideoMAE model, the model achieved high level of performance, indicating the ability to generalize effectively. Figure 2 illustrates training and validation loss over epochs.

The results of this study indicate that transfer learning models are capable of generalizing effectively across a variety of SL classes. It is important to note that the VideoMAE model consistently demonstrated strong performance in the majority of classes, with precision and recall values that varied between 85% and 100%. A few classes, including "fa," "kha", exhibited lower accuracy, which may be identified to their lower representation in the dataset or the inherent difficulty in distinguishing them from others.

TABLE II: Comparison of Model Performance (All Metrics in %)

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**
VideoMAE	95.25	95.71	95.25
TimeSformer	93.44	94.34	93.44
ViViT	92.81	93.34	92.81

The corresponding performance of ViViT, and VideoMAE on the KSU-SSL dataset is shown in Table II. In order to properly recognize complex hand gestures and motions in sign language, VideoMAE's showed a remarkable ability to develop strong spatiotemporal representations from raw video data, as demonstrated by its optimum accuracy. Nonetheless, ViViT generated competitive outcomes, demonstrating its ability to

effectively capture movements and gesture dynamics and efficiently integrates spatial and temporal information. The comparison findings highlight the KSU-SSL dataset's generalizability and robustness across various architectural designs, demonstrating that it is adaptable enough to enable cutting-edge transformer-based models. ViViT's attention mechanism is made it stable for signer-variability, where the KSU-SSL dataset consists of 40 signers, but it computationally expensive, limiting scalability. While TimeSformer's method of split attention is computationally efficient, effective at modeling long-range gesture dependencies, less so for static signs. Both models, pre-trained on Kinetics-400 , transfer well to KSU-SSL's low-resource setting via transfer learning.

V. Discussion

VideoMAE achieves the highest performance on KSU-SSL with an accuracy of 95.25%, 95.71% average precision, and 95.25% average recall, the credit attributed to its masked autoencoding (90–95% masking) that captures robust spatiotemporal features. TimeSformer shows good results as will, achieving 93.44% accuracy, utilizing divided space-time attention for efficient gesture modeling, while ViViT come last with 92.81% accuracy, excelling in signer variability due to factorized attention. The result generalizes well to new and unseen data, this low level of overfitting indicates that this transfer learning strategy has managed to successfully embed an image representation into the pretrained transformer model and extracted relevant features from SL videos. This results show that, the transformer models outperform CNNs (e.g., I3D) by capturing long-term dependencies, which is critical for SSL's dynamic gestures. The effect of signers behavior was observed through the class-wise performance, which considers a strong measure that provide key information of the recognition capability on KSU-SSL dataset. We experimented accuracy, precision and recall metrics and observed patterns related to individual class characteristics that accounts for strengths and weaknesses of the model. An observation showed that the dynamic signs have more average results than static signs, which maybe is due to the short length of static signs, such as "fa" at 62.07% precision, "kha" at 65%, reflecting dataset diversity challenges.

The inclusion of ViViT and TimeSformer in our analysis demonstrates the KSU-SSL dataset's adaptability and robustness. Regardless of different approaches of VideoMAE TimeSformer and ViViT for spatiotemporal video analysis, these models verify the dataset's applicability for a variety of transformer architectures by achieving competitive performance metrics. The models' complimentary strengths are further demonstrated by this comparison. While ViViT's modular framework provide alternate avenues for modeling intricate motion patterns, VideoMAE excels at utilizing pre-trained spatiotemporal features through masked autoencoding. Collectively, these findings show how the KSU-SSL dataset may support a variety of state-of-the-art approaches, hence stimulating innovation in sign language recognition research. The finding additionally has significant implications for future study and real-world applications. The variation of the dataset ensures its applicability to different transformer-based models, leading the way for advancements in sign-based communication systems and accessibility technologies.

a) Future Work: Future work will involve Signer-independent tests to determine the generalizability of the approach across different environments, enabling real-world deployment in a variety of applications. Further study will investigate the efficacy of transfer learning by refining the final classifier on the KSU-SSL dataset using previous layers trained on public sign language datasets like ASL, BSL, or CSL. This method uses common features across many sign languages, which could increase model generalization and recognition accuracy.

VI. Conclusion

This paper presents the fine-tuning of three transformer models for SL recognition task. The paper present a fine-tuning three transformer models that exploit the transfer learning mechanism for SL recognition task. The VideoMAE model achieved 95.25%, which indicate the efficiency of this approach. Moreover, the findings and approaches presented in this paper emphasis on the effeteness of using transfer learning approach to handle the scarcity of large SL resources, specifically for SSL, we advance equitable communication opportunities reaching equal communication opportunities for all individuals, irrespective of their hearing capabilities. Our results indicates that VideoMAE's achieved optimal performance at 95.25% accuracy on isolated SSL classification on KSU-SSL, followed by TimeSformer (93.44%) and ViViT (92.81%), which perform better than CNNs with robust spatiotemporal modeling. Future directions include continuous SSL recognition, transformer extension for mobile deployment, and multimodal and cross-lingual Arabic SLR study, facilitating inclusive communication across the region.

References

[1] M. Alsulaiman, M. Faisal, M. Mekhtiche, M. Bencherif, T. Alrayes, G. Muhammad, H. Mathkour, W. Abdul, Y. Alohali, M. Alqahtani, et al., "Facilitating the communication with deaf people: Building a largest Saudi sign language dataset," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 8, p. 101642, 2023.

[2] S. Al Ahmadi, F. Mohammad, and H. Al Dawsari, "Efficient YOLO-based deep learning model for arabic sign language recognition," *Journal of Disability Research*, vol. 3, no. 4, p. 20240051, 2024.

[3] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 10078–10093.

[4] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," in

Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 6836–6846.

[5] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 813–824.

[6] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert Syst. Appl.*, vol. 164, p. 113794, 2021.

[7] A. A. I. Sidig, H. Luqman, S. Mahmoud, and M. Mohandes, "KArSL: Arabic sign language database," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 20, no. 1, pp. 1–19, 2021.

[8] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2020, pp. 1459–1469.

[9] W. Kay *et al.*, "The Kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[10] H. Hu, W. Zhao, W. Zhou, and H. Li, "Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 11221–11239, 2023.